Reproducibility (climate approach)

Everything you need to know in 10 minutes

xpected a "cold and dry rejection"
xplaining
ed work-
surprised
eason, he
ound for

n replica-
n of luck,
s. Survey
develop-
ing severe
has been
does not
t, he says,
pound to
epted[4].

"REPRODUCIBILITY IS LIKE BRUSHING YOUR TEETH. ONCE YOU LEARN IT, IT BECOMES A HABIT."

than half pointed to insufficien
or low
tion po
reagent
that are
But
by con
develop
Wiscon
and pc
bureau
doing a
stretch
the cost
project.
senior
juniors,
labs wit
and me

heir labs had taken concrete steps
past five years. Rates ranged from
f 24% in physics and engineering

off and make it worse," Kimble s

WHAT CAN BE DONE?

The not-so-good news

EC-Earth3.1 is bugged

The model can only run if catching floating-point exceptions is not enabled(-fpe0 is not enabled)

```
-O2 -g -traceback -vec-report0
```
works

```
-O2 -fp-model precise -fimf-arch-consistency=true
-no-fma -g -traceback -vec-report0 -r8
```
works

```
-O2 -fp-model precise -fimf-arch-consistency=true
-no-fma -g -traceback -vec-report0 -r8 -fpe0
```
Fails at run time

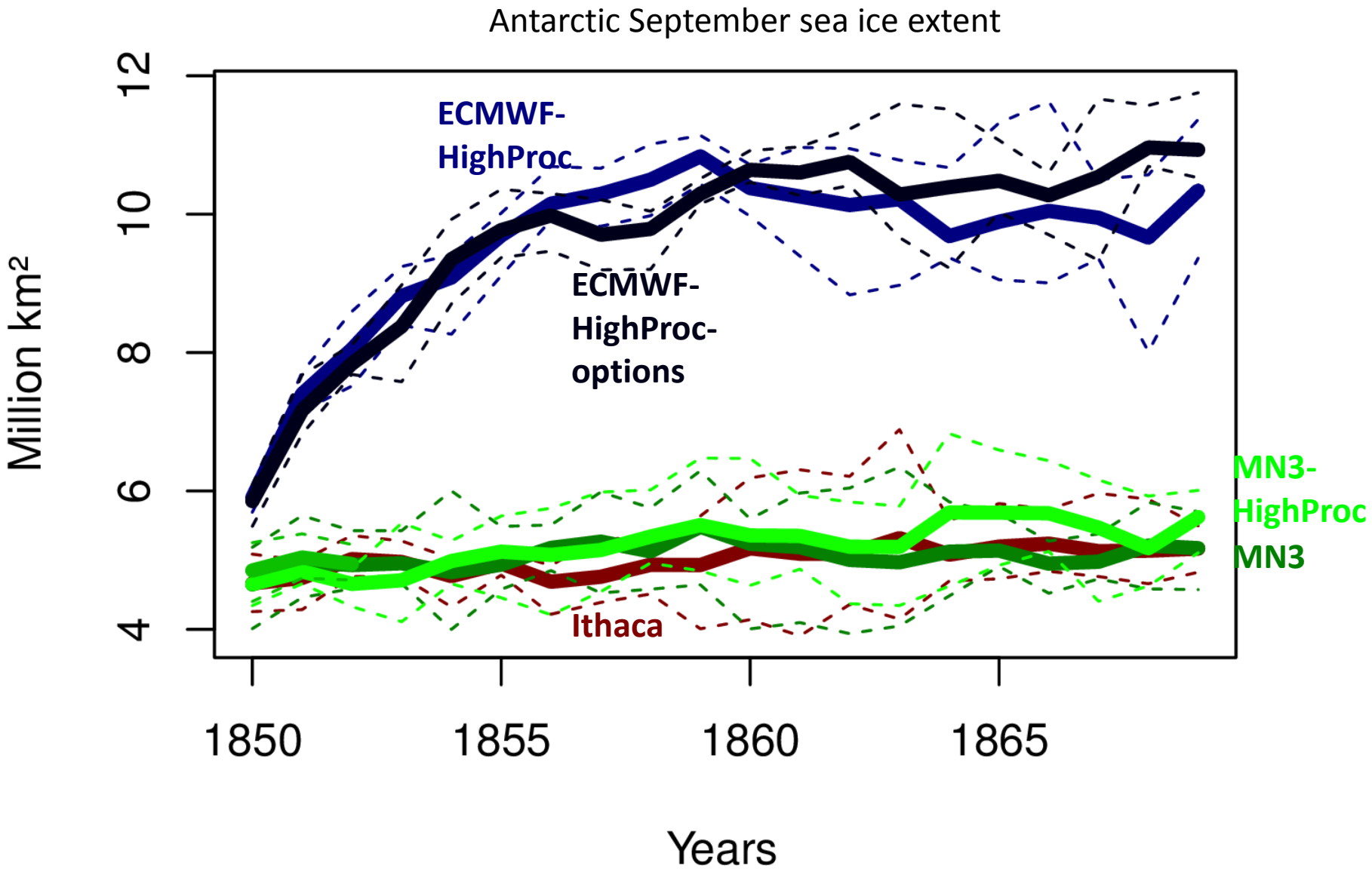
The not-so-good news

EC-Earth3.1 is bugged

The model can only run if catching floating-point exceptions is not enabled(-fpe0 is not enabled)

EC-Earth3.1 is not climate-reproducible under the 3.1 standard configuration

The model has a machine-dependent mean state. One additional experiment is required to make the light.


Antarctic September sea ice extent
Million km²
12
10
8
6
4
1850
1855
1860
1865
Years
ECMWF-
HighProc
ECMWF-
HighProc-
options
Ithaca
MN3-
HighProc
MN3

The not-so-good news

EC-Earth3.1 is bugged

The model can only run if catching floating-point exceptions is not enabled(-fpe0 is not enabled)

EC-Earth3.1 is not climate-reproducible under the 3.1 standard configuration

The model has a machine-dependent mean state. One additional experiment is required to make the light.

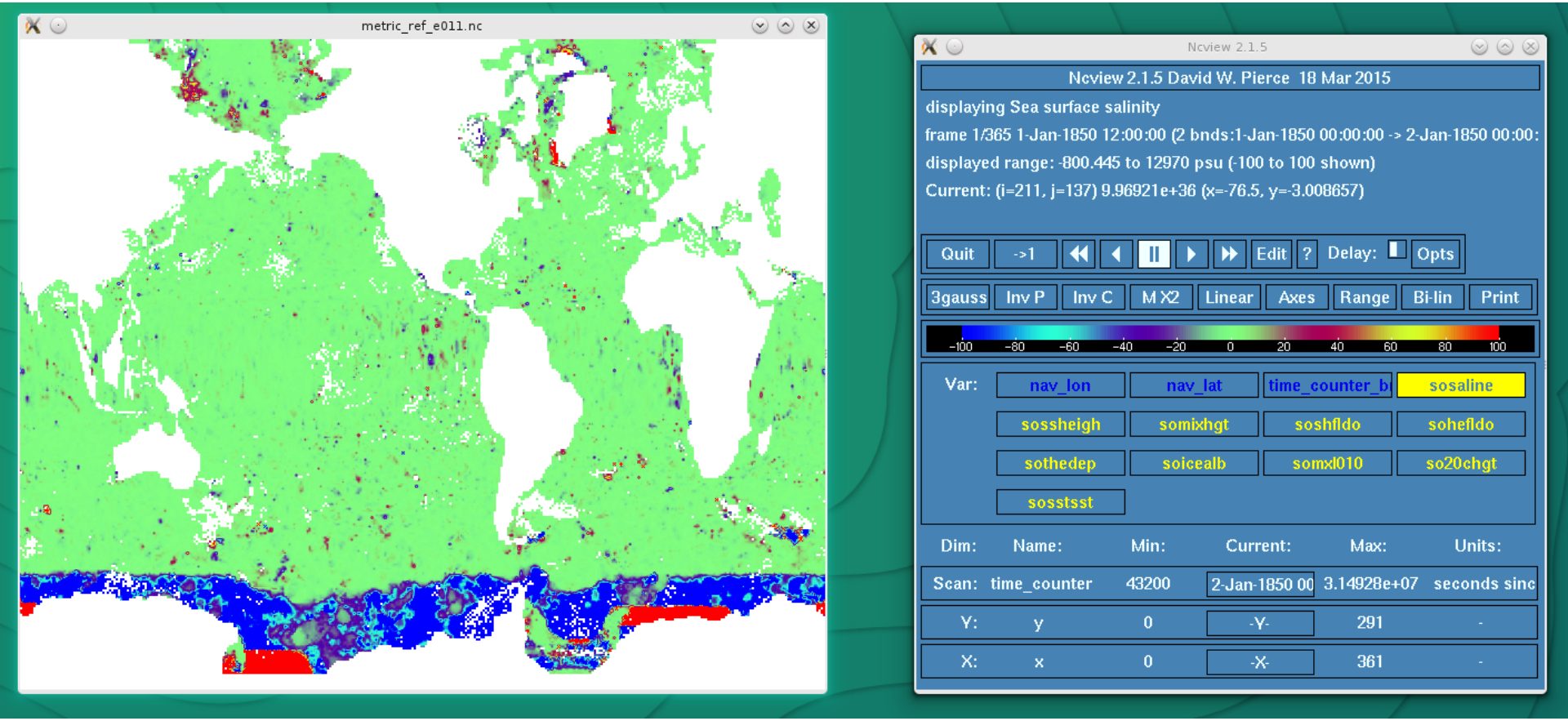
We cannot nail down the cause of non-reproducibility

Perhaps wrong initialization of arrays for river runoffs

1st of January 1850 (1st day of the simulation)

Measure of the ΔSSS as compared to internal variability

$$\frac{\text{mean}(SSS_{ECMWF} - SSS_{MN3})}{\sigma(SSS_{ECMWF})}$$



The good news

We have developed an original and robust protocol

We reached a compromise between temporal & statistical sampling (at least compared to other methods) + metric

The good news

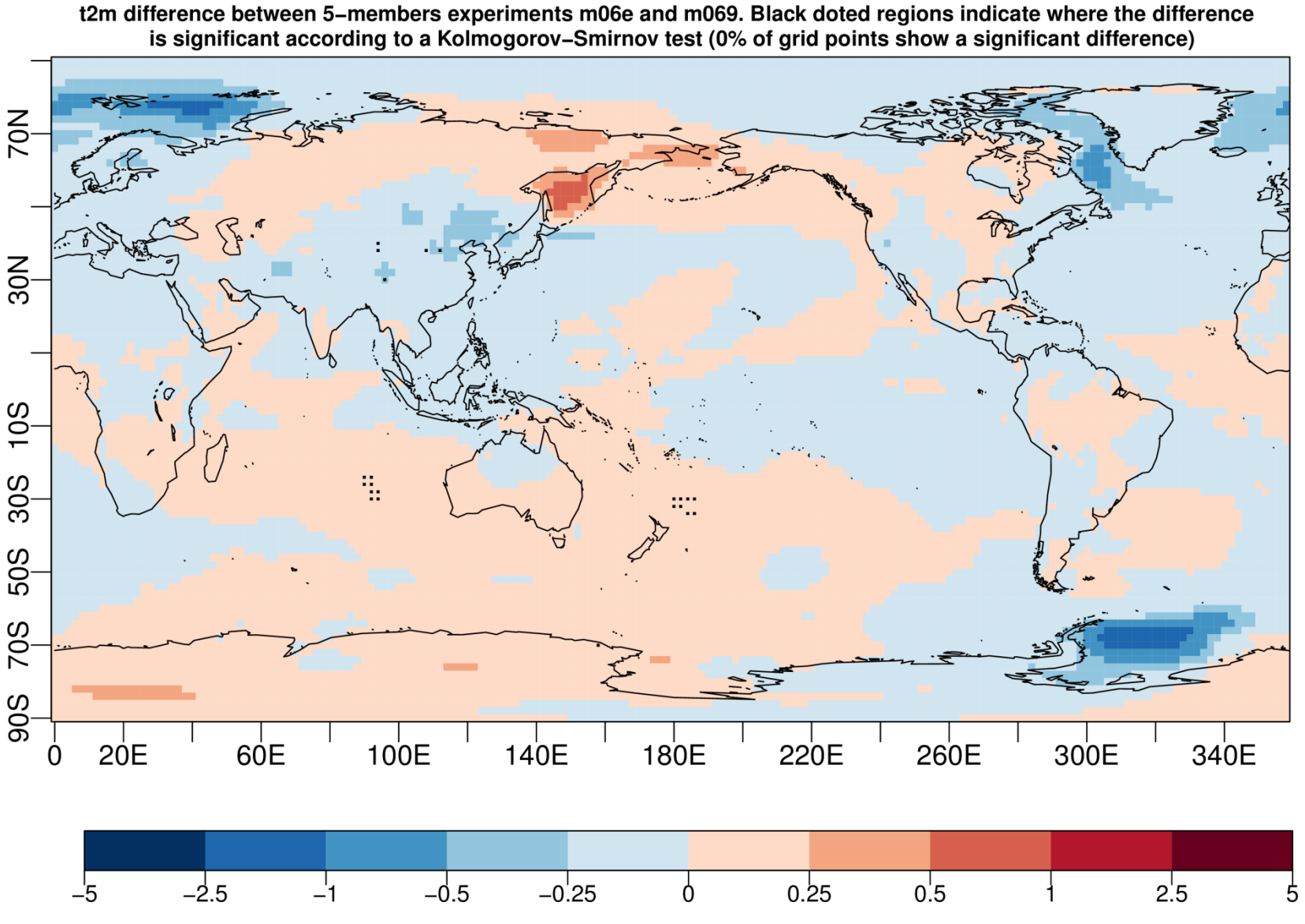
We have developed an original and robust protocol

We reached a compromise between temporal & statistical sampling (at least compared to other methods) + metric

EC-Earth3.1 is clim-reproducible for the #procs

EC-Earth3.1 is clim-reproducible for the compilation options

Changing the number of processors only does not affect the results



The good news

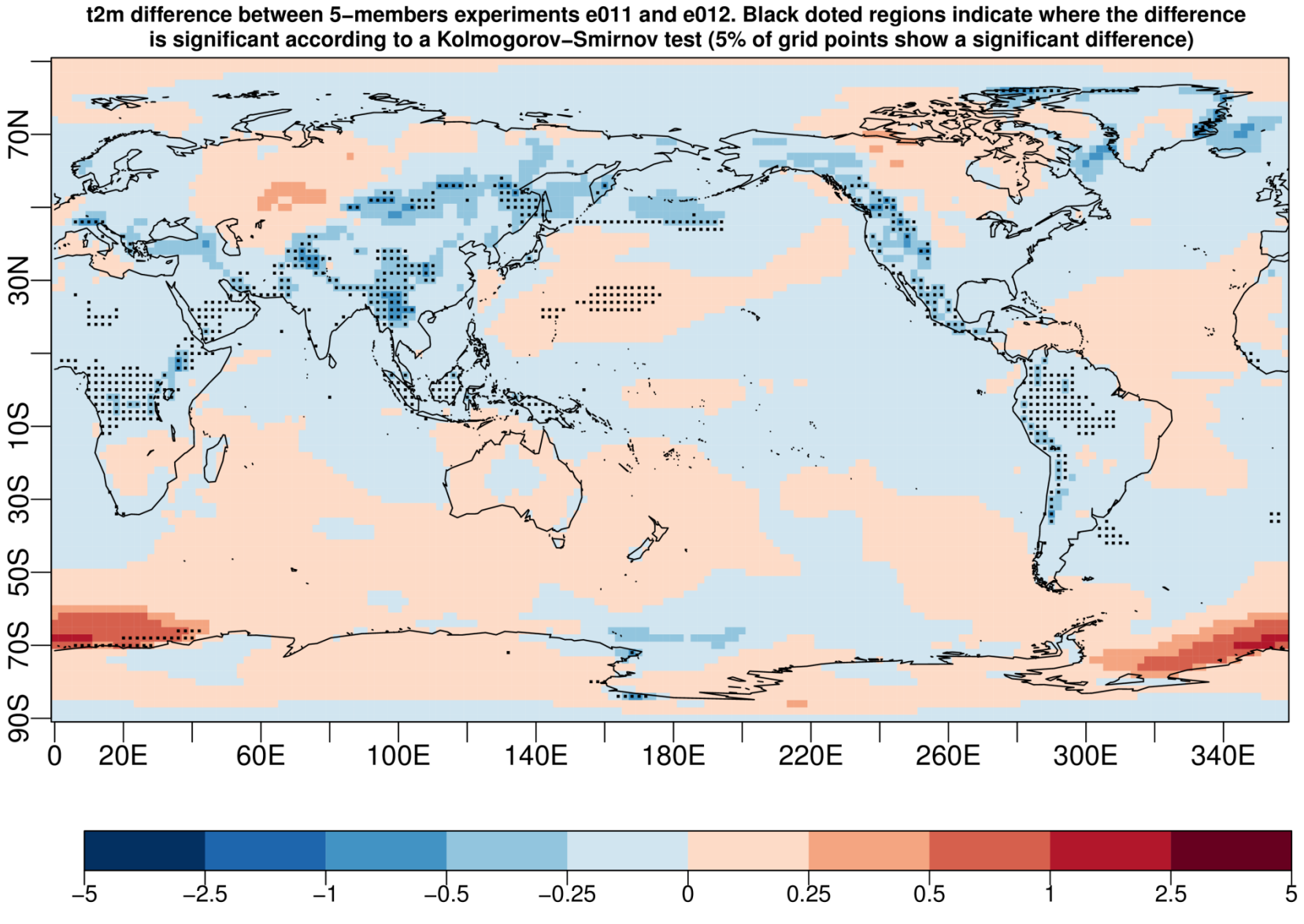
We have developed an original and robust protocol

We reached a compromise between temporal & statistical sampling (at least compared to other methods) + metric

EC-Earth3.1 is clim-reproducible for the #procs

EC-Earth3.1 is clim-reproducible for the compilation options

Changing the compilation options only does not affect the results (but –fpe0 is not enabled in either case)



At this stage we have two options:

1) We run an EC-Earth3.**2beta** experiment with –fpe0 on ECMWF
   a) If it's different from MN3's, then this is a stunning result: even when extra-care is taken about flags, different compilers do provide different results. That's a BAMS/Monthly Weather-type paper, because we did everything we had in our hands to make reproducibility possible, and yet we get different climates. Caveat: we cannot nail down the physical reason for the differences.
   b) If it's the same as MN3's, then it means that we can no longer port codes without activating the –fpe0. Previous results obtained with EC-Earth3.1 and EC-Earth2.3 can be questioned. That's a GMD-style paper.

2) We don't run the extra experiment. In that case we are left with some open questions.

In any case, think about this:
We don't want to bring the discredit on EC-Earth and bite the hand that feeds us. Up to know, irreproducility is **our own fault – we've preferred to ignore warnings.**

What is the key message we want to convey?

1) We have developed **a method** to assess reproducibility, and it is a useful tool to detect when the code is not portable
2) Climate simulations are not reproducible if one does not **pay attention to important details** that are usually meaningless to climate scientists. IT and Climate scientists have to work together.

1) Introduction
    1) Reproducibility is the central concept of exact science
    2) Climate research is no exception; important given high level of interactions
    3) Definitions of bit-reproducibility and climate-reproducibility
2) Methods
    1) How the simulations were conducted, initialized, run. Advantages of long runs and ensemble runs. Git versioning, autosubmit.
    2) How the simulations were analyzed: Reichler and Kim approach (strict)
    3) How the climate-reproducibility was investigated: K-S Smirnoff tests (Omar)
3) Results
    1) Bit-reproducibility on the same machine: yes?
    2) Clim-reproduciblility under processor change: yes
    3) Clim-reproducibility under change of compilation options change: yes
    4) Clim-reproducibility under compiler change: yes if the model is bugged, no else?
4) Discussion
    1) Best practices for CMIP6: the test has to be repeated for every new version.
    2) HPC uncertainty to be added on top of other sources
    3) Necessary for climate scientists to understand the meaning of compilation opts.
5) Conclusion